A single strand-based library preparation method for unbiased virome characterization

Xichuan Zhai [1], Alex Gobbi [2, ‡], Witold Kot [2], Lukasz Krych [1], Dennis Sandris Nielsen [1], Ling Deng [1, #]

[1] Section for Food Microbiology, Gut Health and Fermentation, Department of Food Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark

[2] Section of Microbial Ecology and Biotechnology, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark

‡ Current address: European Food Safety Agency (EFSA), Via Carlo Magno 1A, 43126 Parma, Italy

Corresponding author

Tel.: +45 35330194

E-mail address: lingdeng@food.ku.dk

ABSTRACT: The gut virome is an integral component of the gut microbiome, playing a crucial role in maintaining gut health. However, accurately depicting the entire gut virome is challenging due to the inherent limitations and biases associated with current sequencing library preparation methods. To overcome these problems, we repurposed the ligation-based single-stranded library (SSLR) preparation method for virome studies. We demonstrate that the SSLR method exhibits exceptional efficiency in quantifying viral DNA genomes (both dsDNA and ssDNA) and outperforms existing double-stranded (Nextera) and single-stranded (xGen, MDA+Nextera) library preparation approaches in terms of minimal amplification bias, evenness of coverage, and integrity of assembling viral genomes. The SSLR method can be utilized for the simultaneous library preparation of both DNA and RNA viral genomes. Furthermore, the SSLR method showed its ability to capture highly modified phage genomes, which were often lost using other library preparation approaches.

Keywords: single-stranded library, phage mock community, ssDNA virome, dsDNA virome, RNA virome, gut virome, modification

28 **Main**

29 The gut microbiome constitutes a diverse array of microbes, comprising bacteria, archaea,
30 viruses/bacteriophages, and fungi, residing in the human intestine [1]. Bacteriophages (or phages for short),
31 among them, are believed to be highly abundant constituents of the human microbiome possibly equal in
32 numbers to bacteria. They play important roles in modulating the diversity and abundance of gut bacteria,
33 maintaining a dynamic equilibrium [1, 2]. There is growing evidence linking gut virome dysbiosis to human
34 diseases such as inflammatory bowel disease [3-5], severe acute malnutrition [6], alcoholic liver disease [7], type
35 2 diabetes [8] stunting [9] and recently also asthma [10]. Studies have also demonstrated that in fecal
36 microbiota transplantation to treat e.g. recurrent *Clostridioides difficile* (rCdiff) the virome component is
37 important for treatment efficacy and even that sterile filtered feces used for so-called fecal filtrate
38 transplantation is able to cure rCdiff [11-13]. Consequently, there is growing interest in studying the gut
39 virome to understand its role in various diseases and its potential clinical applications.

40

41 However, this effort is hampered by persistent biases and inconsistencies across studies due to
42 non-standardized and non-optimized pipelines. Every stage of the virome exploration process, from
43 sample collection to bioinformatic analysis, is crucial, with virome characterization of complex microbial
44 communities relying on essential steps such as virus-like particle isolation/purification, viral DNA/RNA
45 purification, and sequencing library preparation and shotgun high-throughput sequencing [14-17].
46 Furthermore, despite advances during recent years, virus identification is still challenging due to their
47 variability, small genome sizes, and genetic mosaicism. Further, virome studies (distinct from bacteriome
48 studies) present a unique challenge with diverse genome types (dsDNA, ssDNA, dsRNA, and ssRNA) and
49 different topologies (linear, circular, or fragments) [18, 19], complicating library preparation and down-stream
50 analysis [1, 3, 20], emphasizing the need for universal or standardized virome sequencing library construction
51 methods for improved reproducibility and comparability across studies.

52

53 Commonly used library preparation methods, such as randomly amplified and linker-amplified shotgun
54 libraries, have limitations, as these methods are restricted to target only dsDNA viruses and can result in
55 uneven coverage of viral genomes, particularly when the viruses are very different in proportional
56 abundances [19, 21]. The transposon-based method, while fast and requiring low-input template, is still
57 restricted to dsDNA [22]. To overcome these limitations, multiple displacement amplification (MDA) has
58 been utilized, increasing DNA amounts for low biomass virome samples and converting ssDNA into dsDNA
59 for further library preparation. However, MDA distorts the ratios of different viral DNA forms by
60 over-amplifying small circular ssDNA genomes and unevenly amplifying linear genomes [23-25]. More
61 recently, single-stranded based methods for library construction have been applied, such as the xGen
62 ssDNA & Low-Input DNA library preparation kit from IDT (previous name Accel-NGS kit from Swift
63 Biosciences). Though promising, the cost per sample is relatively high and it involves multiple purification
64 steps lowering throughput.

65

66 Currently, RNA viruses, especially dsRNA viruses, are significantly undersampled due to the instability of
67 RNA and its incompatibility with common DNA sequencing library preparations [26, 27]. As a result, DNA and
68 RNA viruses are analyzed separately [28]. While recent studies have revealed a greater diversity of
69 environmental RNA viruses than previously thought [26, 29, 30], little information is available on human gut
70 RNA viral communities [31]. The commonly used RNA-Seq technology for investigating RNA viruses requires

71 large amounts of sample material, the preparation process is expensive and time-consuming, and can be
72 further complicated by DNA contamination [32-34].
73
74 To address these challenges, we drew inspiration from the Single Reaction Single-stranded LibrarY (SRSLY)
75 method, originally designed for the library preparation of cell-free DNA and oligo sequencing [35]. We have
76 extended and improved this method for gut virome sequencing (both DNA and RNA) with
77 cost-effectiveness and timesaving in mind. We compared the capability and quantitative accuracy of
78 SRSLY with three different virome library preparation methods (Nextera, MDA amplification and xGen) for
79 Illumina short-read sequencing. We evaluated the types and levels of sequencing bias generated by these
80 protocols using a wide range of DNA/RNA phage mock communities and performed qualitative and
81 quantitative analyses using a diverse mock community with different ratios of 4 different genome types
82 (dsDNA, ssDNA, dsRNA and ssRNA) to further verify the method's applicability for virome studies.
83 Additionally, we assessed reliability, including error rates, composition bias and assembling accuracy of
84 these methods. Finally, we validated the utility and performance of the methods for human fecal virome
85 analysis.
86
87 **Results**
88 **SSLR is an effective method for quantifying DNA phage genomes**
89 We conducted sequencing library preparation using three DNA mock communities (Mock A, B and C;
90 **Fig.S1A** and **S1E**) containing different ratios of dsDNA to ssDNA phages (10%, 50%, and 90% of dsDNA,
91 **TableS1**) with five different methods, namely, Nextera, MDA_0.5h, MDA_1.5h (0.5h and 1.5h referring to
92 the amplification time), xGen and SSLR. Overall, all the phage genomes were detectable even when the
93 total dsDNA inputs were as low as 0.20 ng in the high ssDNA genomes mock (Mock C, **Fig.2A**). However,
94 the efficiency of quantification varied between methods. Nextera library preparation significantly
95 underestimated ssDNA phages (24 to 35 fold, **Fig.2A** and **TableS1**), where only 7 of 9 genomes could be
96 captured (**TableS2**), making it unsuitable for ssDNA genomes studies. Although MDA can be very helpful
97 for ssDNA genome studies, we observed a selective amplification bias of ssDNAs (3-fold, **TableS1**) even
98 with short-time amplification (MDA_0.5h). Notably, no significant amplification biases were observed in
99 mocks with a high percentage of ssDNA genomes (≥ 50%, about 1-fold, **TableS1**). In contrast, both SSLR
100 and xGen accurately recovered the percentage of ssDNA genomes when present in high ratios (≥ 50%,
101 Mock B and C), but slightly underestimated ssDNA genomes when present in low ratio inputs (about 10%,
102 Mock A, **Fig.S2A**).
103
104 Interestingly, we observed quite different quantifications of the T4 genome prepared with different
105 methods (**Fig.2A, 2B, 2D** and **TableS1**). Less than 2% of T4 genomes were sequenced in all the
106 MDA-related libraries while 23% to 34% were reached in the Nextera libraries compared to their
107 corresponding inputs. In contrast, both SSLR and xGen overestimated T4 genomes (1.8 to 3.6 folds, we
108 later found this is due to under-quantification of T4 DNA concentration using Qubit for mock construction,
109 see below), especially in the mock with low T4 genome input (Mock C, **Fig.2A**, **2D** and **TableS1**).
110
111 We further examined the quantification of viral genomes without considering T4 in the mocks (**Fig.2C** and
112 **2D**) and found that the percentage of each genome was consistently accurate across different mock
113 communities when the libraries were constructed by SSLR or xGen. These methods showed strong linear
114 correlations between expected and observed percentage distributions ($p < 2.2e-16$, **Fig.2D**). Genome size

and GC-content of each genome had minimal influence on these quantifications (**Fig.S2B** and **S2C**), and both methods provided a similar overview of DNA phage communities (**Fig.2B**). While the variations could be minimized for the MDA-related methods without including the T4 genome, the quantifications of ssDNA (especially phiX174) were less consistent than the quantification of dsDNA genomes due to their much higher sequencing depths (**Fig.S3** and **TableS1**). The quantification of phages using the Nextera method appeared to be more consistent when the dsDNA phages were calculated exclusively (**Fig.2B** and **2C**).

SSLR has less bias and higher accuracy in sequencing DNA phages compared to existing methods

We then assessed the efficiency and accuracy of the library prepared using different methods with the presence of the T4 genome. The overall mapping rate to the mock genomes from SSLR method was comparable to xGen but higher than Nextera and MDA-related libraries (**Fig.S4A**). Furthermore, SSLR had a lower error rate for the entire mock community compared to Nextera and MDA methods, similar to the rate produced by xGen method (**Fig.S4B** and **S4C**).

When we analyzed individual genomes from SSLR and xGen, and observed mostly even coverage among the genomes, except for sharp spikes near the start and end of each genome (**Fig.S3**). However, the coverage of ssDNA genomes was significantly increased in MDA-related methods and largely neglected by the Nextera method (**Fig.S3** and **Fig.S4D**). Additionally, the effect of GC content on individual genome profiles (GC bias) varied among different library preparation methods. MDA-related and Nextera libraries showed a similar pattern with coverage increasing as GC content increased. However, coverage biases from SSLR and xGen were not apparently associated to their GC content (**Fig.S5**).

Finally, we evaluated the assembly accuracy and efficiency of the DNA phage genomes in the mock community prepared with different library methods. At a same sequencing depth, the assembly from the SSLR prepared library exhibited perfect quality, including the fully reconstructed phage T4 genome (**TableS3**). In contrast, the other library methods provided less integrity with large gaps in the reconstructed genomes. Moreover, the hits of assembled contigs to the reference genomes from the SSLR prepared library showed one contig per reference genome (except for phage P1, where SSLR led to 2 contigs for P1 genome, **TableS4**), where the other methods resulted wide range hits from 0 to 22.

SSLR can quantify highly modified viral DNA

Notably, the T4 genome, in which all cytosines are modified to 5-hydroxymethylcytosine (5-HMC) and further glucosylated (glc-HMC) [36], showed a distinct coverage pattern when prepared with MDA-related methods. The sequencing depths at all genome positions were quite low regardless of the MDA time (**Fig.S3**). This finding prompted further investigation by introducing an unmodified C-sites of T4-c genome, which instead have amber mutations in dCTPase and dHMase genes[37]. Quantification of the T4 genome by Nanodrop showed significantly higher concentration (3.03 - 3.45 folds) than Qubit measurements of the same samples, while the less modified T4-c genome showed less differences (1.65-fold, **Fig.3A**). MDA amplification of T4 genome resulted in relatively lower amount of product with shorter fragments compared to phage genomes P1, T7 and T4-c (**Fig.3B** and **3C**). Equal amounts of phage genome (11.11% of each, quantified by Nanodrop) was used for preparation of Mock G (with T4) and Mock H (with T4-c), we observed that T4-c still yield lower percentage after sequencing (**Fig.S6**), but had higher ratios (about 2-fold) than T4 in Nextera, SSLR and xGen method (**Fig.S1C**, **Fig.S1G**, **Fig.3D** and **TableS5**).

159

160 **SSLR can simultaneously identify DNA and RNA viruses**

161 Next, we applied the SSLR method for simultaneous sequencing of DNA and RNA phage genomes (**Table1**,
162 **Fig.S1B** and **S1F**). We used DMSO or heat treatment to denature the dsRNA to facilitate the reverse
163 transcription (RT) process. We found both treatments including the control (No DMSO) could increase the
164 ratio of ssRNA compared Heat or No Heat, especially for the low ssRNA mock (5% of ssRNA, Mock D)
165 (**Fig.4A** and **Fig.S7**). Although pretreatment with DMSO increased the number of reads of dsRNA
166 compared to non-DMSO treatment in the mocks with less dsRNA (5% ~ 25% of dsRNA, Mock D and E,
167 **TableS6**), it was not as effective as heat treatment in terms of read accuracy and sequence coverage
168 (**Fig.4B**, **Fig.S7** and **TableS6**).

169

170 **SSLR provides high quality fecal virome characterization and allows detection of highly modified phage**
171 **genome**

172 Finally, we tested the capability for determining fecal virome composition using the methods described
173 above (**Fig.S1D**). Although both MDA based methods yielded more reads that can be assigned to
174 Reference Viral Database (RVDB) (range, 5.77-6.26%, **Fig.S8A** and **TableS7**) compared to SSLR (0.83-1.20%)
175 and xGen (1.23-1.38%) (**Fig.S8B**), 96% of viral operational taxonomic units (vOTUs) were shared among
176 different methods (**Fig.S8C**). We also observed that although the MDA method resulted in higher
177 alignment rates, it led to lower virome Shannon diversity compared to the other library preparation
178 methods (**Fig.S8D** and **8E**). And for both short and long-time MDA, proportions of ssDNA viral families
179 were increased compared to other library preparation methods (**TableS8**). The SSLR could capture around
180 2-times less of single genome virome (ssDNA of *Microviridae* and ssRNA of *Virgaviridae*) than xGen
181 method (**TableS8**, **Fig.5B** and **Fig.S8F**).

182

183 Additionally, the different library preparation methods led to differential abundant core species as
184 compared with the Nextera method although the majority of vOTUs are not identified at family level
185 (**Fig.S8F**). Based on the virome abundance (log10(RPKM+1)), 4 clusters were identified (**Fig.S9A**). Notably,
186 the abundance of cluster 3 from the SSLR method was approximately 10^3 times higher than the other
187 methods (**Fig.5D** and **Fig.S9B**). We found that these 34 vOTUs were associated with multiple hits to
188 Sphingomonas phage PAU (previous *Myoviridae*) (**Fig.5C, Fig.5D** and **Fig.S9C**). As it is evident that the
189 cytosines in Sphingomonas phage PAU genome were highly modified [38], this further confirmed that SSLR
190 method can efficiently include modified viral genome in the sequencing libraries.

191 **Discussion**

192 The viral community in the human gut is highly complex and likely plays an important role in maintaining
193 gut homeostasis [1, 39]. However, whether the true diversity and distribution of gut virome can be accurately
194 determined by traditional nucleic acids library construction and sequencing methods and with what kind
195 and degree of biases are only partly known. In this study, we addressed these problems by introducing a
196 series of phage genome mock communities reflecting different ratios of four types of viral genomes
197 (dsDNA, ssDNA, dsRNA, ssRNA). To ensure accuracy and avoid extraction efficiency variations [40, 41], we
198 used phage genome (DNA/RNA quantified by Nanodrop or Qubit) as initial inputs instead of phage
199 particles by plaque assay or SYBR green counts. The single-stranded library preparation method (SSLR) we
200 introduced for gut virome studies proved to be highly accurate, less biased and comparably efficient, with
201 the added benefits of being fast and cost-efficient.

202

203 SSLR demonstrated excellent efficiency in quantifying both dsDNA and ssDNA genomes, overcoming the
204 major limitation of the Nextera protocol, which only works for dsDNA genomes [7, 42]. While MDA can be
205 applied for turning ssDNA into dsDNA before using Nextera, and it is known for its amplification
206 preference for ssDNA genomes [23-25]. In line with the previous studies [24], we showed that MDA led to
207 2-3-fold overestimation of ssDNA in low ssDNA mocks (Mock A, around 10% of ssDNA) but not in high
208 ssDNA genome mocks (≥ 50%, Mock B and C). Additionally, the amplification bias appeared to depend on
209 the type of ssDNA genome, as MDA preferentially amplified the alpha3 genome more than PhiX174 and
210 M13mp18, which might be associated to the circular ssDNA of alpha3 being amplified faster than linear
211 ssDNA or the different responses of ssDNA genomes to the heat-denaturing before MDA [23]. Similar to
212 previous studies we also found that SSLR and xGen slightly underestimated ssDNA in mocks with a low
213 ratio of ssDNA (≤ 10%), this might be due to the ssDNA being more sensitive to shearing during the
214 sonication step compared to dsDNA [24, 43].

215

216 T4 is an exception among dsDNA genomes in the mocks, as its relative abundance in SSLR and xGen was
217 consistently higher than its input (1.6 - 3.7 fold) and higher than its abundances in Nextera (1.96 - 9.59
218 fold) and MDA (99 - 586 folds). Interestingly, we observed a significant and consistent difference between
219 the quantification of DNA concentration by Qubit and Nanodrop when measuring T4 genomes. This
220 discrepancy may be attributed to glucosylated 5-hydroxymethyl-cytosine (glu-5HMC) modifications
221 present in the T4 genome [36, 37]. These modifications could interfere with or alter the binding modes of the
222 Qubit dye, leading to lower concentration values than those obtained through UV absorbance-based
223 quantification (Nanodrop) [44]. Considering the high quality of our T4 genome, which was extracted,
224 purified, and recovered from agarose gel (see the method part), the concentration values measured from
225 Nanodrop should be more accurate, as 3 times concentration higher from nanodrop is corresponding to
226 the ~3-times higher sequencing output (**Fig.3A** vs **Fig.2A**). Our results also shed light on the reason why
227 the 7-deazaguanine modified Cellulophaga phage phi38:2 (previous *Myoviridae*) was overestimated in the
228 A-LA method (xGen) and underestimated in MDA [24], as it was demonstrated recently that this phage has
229 extensive modifications on its genome to protect its DNA from bacterial defense systems [45]. It was also
230 reported that when measuring potentially highly modified viral DNA using DNA-binding fluorescent dyes,
231 greater caution should be exercised to avoid pitfalls [46]. In comparison to the T4 genome, the T4-c genome
232 is less modified, and its genome can be sequenced deeper using the Nextera method (**Fig.S6**). However, in
233 the MDA-related method, the genome yield was not increased for T4-c suggesting the remaining
234 modifications in T4-c genome could still inhibit the MDA process on a similar level.

235
236 SSLR has an advantage for sequencing highly modified viral genomes (such as T4 with 5-HMC) compared
237 to existing methods, which is further demonstrated by the detection of PAU phage in the fecal virome. We
238 failed to identify PAU-like phage from four gut virome databases (GVD, GPD, MGV and IMG_VR4.1,
239 **TableS10**) [47-50], probably due to the modified phages were not included in the original metagenomic
240 sequencing libraries. This finding suggested that many virome studies may have missed or at least
241 underestimated the existence of viruses with highly modified genomes.
242
243 Unlike the widely used Nextera library preparation method, which can be affected by genome size and
244 GC-content [51, 52], the proposed SSLR protocol exhibits low read errors and provides comprehensive ability
245 for variable genomes with exceptional evenness of coverage and near-complete assembly of phage
246 genomes, as also exemplified for the otherwise "difficult" phage T4. We also observed that SSLR
247 outperforms xGen in several performance metrics (including assembly quality, evenness of genome
248 coverage and error rate, (**TableS4** and **Fig.S3**) probably due to the fact that it does not need synthesizing a
249 second strand for ssDNA viral genomes prior to adaptor ligation, thus retaining the native termini and
250 avoiding possible artifacts or errors presenting during sequencing library preparation [35].
251
252 SSLR also offers the advantage of using cDNA directly, the first-strand products of RT, as ligation templates
253 for sequencing of RNA viruses. This simplifies the RT step and preserves the strand information [35]. Our
254 DNA and RNA genome mocks tests demonstrated that the RT step did not significantly affect the
255 distribution of DNA viruses, although the ssRNA virus MS2 was overrepresented. This overrepresentation
256 can be attributed to the faster RT conversion rate of ssRNA compared to dsRNA, leading to non-uniformity
257 in read coverage, which is consistent with previous observations [53]. Furthermore, for denaturing of dsRNA,
258 the heat-shock treatment appears to be a more cost-effective and easier approach compared to DMSO
259 treatment, it not only reduces the denaturation time but also eliminates the necessity to remove DMSO
260 after treatment, thereby conserving more sample [53].
261
262 In conclusion, we introduce an improved, rapid, straightforward, and efficient ligation-based
263 single-stranded DNA library preparation method tailored for virome studies.

264 **Materials and Methods**

265 **Mock communities with phage DNA and RNA genomes**

266 Customized bacteriophage mock community samples were prepared, comprising of a mixture of 4 to 9
267 bacteriophage genomes with different genome sizes (3.5 to 168.9 kbp), genome types (dsDNA, ssDNA,
268 dsRNA and ssRNA), and G+C content (35 to 56%) (**Table1**). The phages were propagated using the
269 conditions listed (**TableS10**) and the genomic DNA/RNA extractions were conducted using commercial
270 extraction kits (DNeasy Blood & Tissue Kit for DNA or RNeasy Mini Kit for RNA, both from Qiagen) and
271 purified with GeneJET Genomic DNA Purification Kit (Thermo Scientific, #K0722) for recover the
272 high-quality genome. The quality and integrity of DNA and RNA was checked by agarose gel
273 electrophoresis and Agilent TapeStation 4200 High Sensitivity (HS) RNA ScreenTape. Their concentrations
274 were quantified by dsDNA (or ssDNA) HS Assay Kit or RNA HS assay kit (ThermoFisher Scientific, Waltham,
275 MA, USA) with Qubit4.0, along with Nanodrop1000 (V3.8.1).
276

277 **DNA phages library preparation and sequencing**

278 For the DNA phage library preparation, we mixed genomic DNA of 9 phages (7 dsDNAs and 2 ssDNAs)
279 with different ratios (Mock A, Mock B and Mock C) to test 5 different library preparation methods
280 (**Fig.S1A, Fig.S1E** and **TableS1**).
281

282 Nextera XT libraries (Nextera) were prepared according to the manufacturer's protocol (FC-131-1096)
283 (Illumina). Briefly, 1.0 ng of each DNA library was enzymatically fragmented and tagged by tagmentation.
284 Amplification was performed using Illumina dual index (i7 + i5) adapters and cleaned by AMPure XP bead
285 cleanup (A63880l; Beckman Coulter). Two technical replicates were generated for each biological sample,
286 resulting in 6 Nextera-prepared libraries.
287

288 The same mixture of phage DNA as mentioned above was amplified by Multiple Displacement
289 Amplification (MDA) using the Genomephi V3 kit (GE Healthcare Life Science, Marlborough, MA, USA)
290 with 10 ng input and incubated at 30°C for 0.5 h (MDA_0.5h) or 1.5 h (MDA_1.5h). Subsequently, the
291 amplified DNA was cleaned using Genomic DNA Clean & Concentrator™ kit (Zymo Research, Irvine, CA,
292 USA). Qubit® 1X dsDNA HS Assay Kit on Qubit Fluorometer (Life Technologies, CA, USA) was used to measure
293 the concentration of the purified DNA. Finally, the library was constructed using the Nextera XT kit
294 (Illumina, San Diego, CA, USA). Two technical replicates were generated for each biological sample,
295 resulting in a total of 12 MDA-prepared libraries.
296

297 xGen™ ssDNA & Low-Input DNA Library (xGen, previously known as Swift Bioscience Accel-NGS® 1S) was
298 prepared following the manufacturer's protocol (Catalog # 10009859, Integrated DNA Technologies, IDT).
299 The Mock (15 ng in 20 µl TE buffer) was randomly fragmented by BioRuptor (Diagenode, Liege, Belgium)
300 at default intensity for 15 sec on and 90 sec off, with 7 cycles at 4°C as described by Hoang et al [55].
301 Sheared DNAs were firstly subjected to 3' end tailing and ligation with adapter1, followed by extension
302 and ligation with adapter2. The purified ligation products were indexed with xGen™ Unique Dual Index
303 (UDI) Primers (Catalog # 10005975, IDT) and amplified through a 10-cycle PCR. The amplified products
304 were cleaned up with AMPure XP beads (Beckman Coulter Genomic, CA, USA). For each biological sample,
305 two technical replicates were conducted, resulting in a total of 6 xGen-prepared libraries.
306

307 For the Single-stranded library preparation (SSLR, detailed in **Supplementary file1**), the same mocks were
308 used. After random fragmentation, the end modified adapters (**Supplementary file2**) synthesized by IDT
309 (IDT, Leuven, Belgium) were ligated to DNA fragments employing the method from Troll et al [35].
310 Subsequently, the ligation mix was purified using MinElute Reaction Kit (Qiagen, Hilden, Germany) and
311 the ligation products were eluted with 23 µl of TE buffer (10 mM, pH8.0). The cleaned ligation products
312 were then amplified and indexed for sequencing with Illumina dual Indexes (i7 + i5) and further purified
313 using AMPure XP beads (Beckman Coulter Genomic, CA, USA). An additional bead cleanup step can be
314 applied to mitigate the formation of byproducts and adapter dimers resulting from self-ligation of the
315 adapters (**Fig.S10**). Two technical replicates were performed for each biological sample, resulting in a total
316 of 6 SSLR-prepared libraries. A detailed protocol can be found on (**Supplementary file1**).
317
318 The DNA concentrations of Nextera-prepared, MDA-prepared, xGen-prepared and SSLR-prepared libraries
319 were measured by Qubit® 1X dsDNA HS Assay Kit on Qubit Fluorometer (Life Technologies, CA, USA) and
320 fragment length distributions were determined using TapeStation 4200 (Agilent, CA, USA). Equal amounts
321 of DNA from each library (except xGen-prepared) were pooled and sequenced using 2×150 bp paired-end
322 settings on an Illumina NextSeq550 platform (Illumina, CA, USA). The xGen-prepared libraries were
323 sequenced separately on an iSeq100 System (Illumina, CA, USA) using 2×150 bp paired-end settings due
324 to their indexes being incompatible with the Illumina index.
325
326 **RNA phages library preparation and sequencing**
327 Four types of phage genome (**Table1,** dsDNA, ssDNA, dsRNA and ssRNA) with different ratios were
328 prepared for phage DNA/RNA mock communities (Mock D, Mock E and Mock F). We used three mock
329 communities to determine the applicability of the SSLR method for quantification of RNA and DNA phages
330 simultaneously (**Table1, Fig.S1B** and **Fig.S1F**). Specifically, 20 µl mock community was treated with DMSO
331 at a final concentration of 50% and incubated at 65°C for 90 min [53]. The DMSO was subsequently
332 removed using a QIAmp viral RNA mini kit (Qiagen, Hilden, Germany), according to the manufacturer's
333 instructions. As a control, a DMSO-free treatment was included. Alternatively, the same mock community
334 was denatured at 95°C for 3 min and immediately snap-cooled on ice, maintaining the nucleic acids as
335 single-stranded. Again, a sample without heat treatment was included as a control library.
336
337 All the DMSO-treated samples, heat-treated samples and their corresponding controls were subjected to
338 reverse transcription in a 20-µl reaction system according to the user guide (SuperScript™ IV VILO™
339 Master Mix, Invitrogen™) and the reaction mix was purified using DNeasy Blood & Tissue Kits (Qiagen,
340 Hilden, Germany) after RT with modifications, where the purification procedure started from step 3,
341 whereby an equal volume of ethanol (96%) was added to the PCR reaction, and a 20 µl of Tris buffer (10
342 mM, pH8.0) was used for the final elution. Finally, the library preparations with SSLR were performed
343 according to the above-mentioned steps and sequenced on a NovaSeq6000 platform (Illumina, CA, USA).
344
345 **Sequencing of highly modified phages**
346 Two types of DNA phage mock communities (Mock G and Mock H) were prepared with equal input (11.11%
347 for each of the 9 genomes, **Fig.S1C** and **Fig.S1G**). Mock G contained a highly modified T4 genome (T4),
348 while Mock H contained a less modified T4-c genome (T4-c). All the libraries were prepared as described
349 above and sequenced with the NovaSeq6000 platform (Illumina, CA, USA).
350

351 **Human fecal virome isolation and sequencing library preparation**

352 For environmental sample, a fresh fecal sample was obtained from an anonymous healthy adult (Ethical
353 Committee of the Capital Region of Denmark registration number H-20028549) and thoroughly mixed
354 with SM buffer. Fecal virome isolation and purification were carried out according to our previous method
355 with the following modifications [16]: the Centriprep 50K was replaced by Centrisart® I centrifugal
356 ultrafiltration unit (MWCO 100 kDa, Sartorius Stedim Biotech GmbH) and the enrichment step was
357 conducted at 2500 × g for 30 min at 4°C. The QIAmp viral RNA mini kit (Qiagen, Hilden, Germany) was
358 used for the extraction of viral DNA/RNA from the concentrated virome solution (**Fig.S1D**). The library
359 preparations were conducted with the aforementioned methods in triplicate. An equal amount of DNA
360 from each library was pooled and sequenced using 150 bp paired-end settings on an NextSeq550
361 platform (Illumina, CA, USA).

362

363 **Metavirome sequencing, data pre-processing and data analyses**

364 Metavirome reads were quality filtered, trimmed and assembled using a previously published pipeline
365 from GitHub: https://github.com/jcame/virome_analysis FOOD. Detailed data analyses for sequencing of
366 the mock communities and fecal virome can be found from **Supplementary file3**. For mock community
367 data, plots were generated using the ggplot2 [56] package in R (v4.2.0). Analyses of viral community α- and
368 β-diversity were performed using packages Phyloseq (v1. 36.0) [57] and Vegan (v2.6.2) [58] in R. For α-diversity
369 analyses, indexes of observed taxa and Shannon diversity were calculated with t.test using the package
370 ggsignif (v0.6.3) [59]. For β-diversity analyses, Bray-Curtis distance metrics were calculated and
371 unconstrained ordination was performed using principal coordinate analysis (PCoA). To identify
372 differentially enriched vira on the summarized family level, DESeq2 (v1.42.0) was adopted [60]. The results
373 were then visualized in a heatmap using the R package complexheatmap (v2.18.0) [61]. Estimation of
374 sequencing error rates was conducted according to Wilcox et al [53]. Coverage biases were visualized using
375 the python script from: https://github.com/padbr/gcbias, with modifications made by adding the Pearson
376 correlation. To visualize the functional genes of annotated viral contigs, DnaFeaturesViewer (v3.1.0) [62] and
377 R package gggenomes (v 0.9.12.9000) [63] was used.

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplemental information files. All the raw viral metagenome sequences data produced in this study are available through the National Center for Biotechnology Information's Sequence Read Archive under BioProject accession number PRJNA1094595 with submission No. SUB13375064 (DNA mocks), No. SUB14350434 (DNA/RNA mock), No. SUB14350532 (Modification mock), and No. SUB14350824 (fecal virome).

The public databases for checkV (v1.5) at https://portal.nersc.gov/CheckV, VIBRANT (v1.2.1) at (https://github.com/AnantharamanLab/VIBRANT/tree/master/databases), Virsoter2 at (https://osf.io/u3t4j), VOG217 at (https://fileshare.csb.univie.ac.at/vog/vog217) and Iphop (v1.3.2) at (https://portal.nersc.gov/cfs/m342/iphop/db/) are accessible online.

Code availability

This paper does not generate the original R code for data analysis and visualization.
Code for metavirome pre-analysis is available from GitHub: https://github.com/jcame/virome_analysis-FOOD.
Original python code from (https://github.com/padbr/gcbias) with modifications for fitting the GC bias can be found from GitHub: https://github.com/padbr/gcbias.
The read accuracy and mutation analysis can be found on GitHub: https://github.com/awilcox83/dsRNA-sequencing.
Further information and requests for data, code and resources required to reanalyze the data reported in this paper should be directed to and will be fulfilled by XC. Zhai and L. Deng.

Author information

Authors and Affiliations

Section for Food Microbiology, Gut Health and Fermentation, Department of Food Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark

Xichuan Zhai, Lukasz Krych, Dennis Sandris Nielsen, Ling Deng

Section of Microbial Ecology and Biotechnology, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark

Alex Gobbi, Witold Kot

European Food Safety Agency (EFSA), Via Carlo Magno 1A, 43126 Parma, Italy

Alex Gobbi

Contributions

This research was designed and directed by LK, DSN and LD. The experiment was performed by XZ. Data collection was performed by XZ, AG and WK. Data analysis was done by XZ and LD. The manuscript was written and revised by XZ, DSN and LD. All authors read and approved the final manuscript.

Corresponding authors

Correspondence to Ling Deng, lingdeng@food.ku.dk

Additional data

Additional file 1: SSLR adapter design.
Additional file 2: Step-by-step library preparation protocol for the reproposed SSLR.
Additional file 3: Metavirome sequencing and data pre-processing.
Additional file 4: Supplementary Table S1-S10. **TableS1**: Summary of DNA genome sequences. Top table: Relative abundance and coverage of phage genomes in libraries from DNA mock communities (Mock A, Mock B and Mock C) with T4 genome. The relative abundance of each genome was calculated based on its coverage by using bowtie2 alignment with our customed database (**Additional file6)**. For comparing the efficiency of different methods for each group of phages (i.e., dsDNA and ssDNA), relative abundance of two type genomes were computed from this table (by dividing the value for each phage by the sum for the group). Middle table: The average coverage for each genome from each library preparation method. Bottom table: similar to top table without T4 genome. **TableS2**: Comparisons of cost and effectiveness of

Mock virome library preparation method. This comparison only considers the library preparation step for DNA virome. **TableS3**: Overview of metavirome sequencing for the DNA phage mock community. Statistics from assembly before and after quality checks with CheckV, vibrant and virsorter2. Contigs from each library prepared methods were evaluated by quality checks with CheckV, vibrant and virsorter2 and then subjected to the Quast for quality assessment of assembled contigs. The theatrical statistics of mock community were also subjected to Quast and bolded at the bottom of table. **TableS4**: Summary of contigs hits to the customed mock database. The quality-checked contigs were subjected to the customed database and the hit number was counted and listed in the table. **TableS5**: Summary of DNA genome sequences with interested modified genomes (T4 and T4-c). T4 is a highly modified genome, T4-c has lower modification compared to T4 genome. **TableS6**: Summary of DNA/RNA genome sequences prepared with SSLR method. Relative abundance and coverage of phage genomes in libraries from DNA/RNA mock communities (Mock D, Mock E and Mock F) with different treatments (heat, no heat, DMSO and no DMSO). The relative abundance of each genome was calculated based on its coverage by using bowtie2 alignment with our customed database. For comparing the efficiency of different treatments for each type of phage genome (DNA and RNA), relative abundance of two type genomes (DNA or RNA) were calculated from this table (by dividing the value for each phage by the sum for the genome types). Middle table: The average coverage for each genome from each library preparation method. The average coverage for each genome from each treatment method is listed in the bottom table. **TableS7**: Taxonomic classification of fecal metavirome sequencing reads into the categories viral, human, bacterial, and unknown origin. To check the presence of non-viral DNA sequences, 50,000 random forward reads were used according to their match to a range of viral, bacterial, and human reference database of Kaiju 1.8.2. **TableS8**: Relative abundance of fecal virome prepared with 5 different library strategies at taxonomy of family level (support data for **Fig.5B**) and the overall alignment rate of different library preparation methods based on bowtie2 alignment to the vog217 database. **TableS9**: Top 5 hits of PAU phage on 4 virus databases (GVD, GPD, MGV and IMG_VR4.1). **TableS10**: Characteristics of phage genomes included in the mock communities from the present study, as well as growth conditions for the strains in the mock communities. **TableS11**: Average alignment rate of each library to the customed databases for all the mock communities tested in the present study.

Additional file 5: Supplementary figures S1-S10.

Funding

This work was supported by research grants (23145 and 36242) from VILLUM FONDEN.

Acknowledgements

We thank the donors and our colleagues at Section for Food Microbiology, Gut Health and Fermentation, Department of Food Science, University of Copenhagen (KU FOOD) for their participation and cooperation. We acknowledge Dr. Samuel Kilcher and Prof. Martin Loessner from Institute of Food, Nutrition, and Health, ETH Zurich for providing phage P35 and its host strain *Listeria monocytogenes* Mack. We acknowledge Dr. Yuvaraj Bhoobalan-Chitty from Department of Biology, University of Copenhagen (Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark) for providing phage T4-c. XZ was supported by China Scholarship Council (CSC) Grant #201906870027.

Abbreviations

SSLR : single-stranded library; Single Reaction Single-stranded LibrarY: SRSLY; RT: Reverse transcription; MDA: multiple Displacement Amplification; PCoA: principal coordinates analysis; vOTUs: viral-operational taxonomic units; ORFs: open reading frames; NCLDV: nucleocytoplasmic large DNA viruses; LCA: Lowest Common Ancestor; SGBs: species-level genome bins; NS: not significant; gut virome database: GVD; gut phage database: GPD; metagenomic gut virus: MGV; high- Integrated Microbial Genomes /Viruses 4.1: IMG/VR4.1; High Sensitivity: HS; Unique Dual Index: UDI; Integrated DNA Technologies: IDT

Ethics approval and consent to participate

A healthy adult anonymous donor donated fecal samples for the study after being informed orally and in writing and provided written consent (Ethical Committee of the Capital Region of Denmark registration number H-20028549).

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Consent for publication

Not applicable.

494 **References**

495 1. Shkoporov, A.N. & Hill, C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell*
496 *Host & Microbe* **25**, 195-209 (2019).
497 2. Van Belleghem, J.D., Dąbrowska, K., Vaneechoutte, M., Barr, J.J. & Bollyky, P.L. Interactions between
498 Bacteriophage, Bacteria, and the Mammalian Immune System. *Viruses* **11**, 10 (2019).
499 3. Liang, G., Cobian-Guemes, A.G., Albenberg, L. & Bushman, F. The gut virome in inflammatory bowel diseases.
500 *Curr Opin Virol* **51**, 190-198 (2021).
501 4. Norman, J.M. et al. Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. *Cell*
502 **160**, 447-460 (2015).
503 5. Fernandes, M.A. et al. Enteric Virome and Bacterial Microbiota in Children With Ulcerative Colitis and Crohn
504 Disease. *J Pediatr Gastroenterol Nutr* **68**, 30-36 (2019).
505 6. Reyesa, A. et al. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *P Natl Acad*
506 *Sci USA* **112**, 11941-11946 (2015).
507 7. Duan, Y. et al. Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease. *Nature* **575**,
508 505-511 (2019).
509 8. Rasmussen, T.S. et al. Faecal virome transplantation decreases symptoms of type 2 diabetes and obesity in a
510 murine model. *Gut* **69**, 2122-2130 (2020).
511 9. Mirzaei, M.K. et al. Bacteriophages Isolated from Stunted Children Can Regulate Gut Bacterial Communities
512 in an Age-Specific Manner. *Cell Host Microbe* **27**, 199-212.e5 (2020).
513 10. Leal Rodríguez, C. et al. The infant gut virome is associated with preschool asthma risk independently of
514 bacteria. *Nat Med* **30**, 138-148 (2024)
515 11. Ott, S.J. et al. Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile
516 Infection. *Gastroenterology* **152**, 799-811 e7(2017).
517 12. Kortright, K.E., Chan, B.K., Koff, J.L. & Turner, P.E. Phage Therapy: A Renewed Approach to Combat
518 Antibiotic-Resistant Bacteria. *Cell Host Microbe* **25**, 219-232 (2019).
519 13. Zuo, T. et al. Bacteriophage transfer during faecal microbiota transplantation in infection is associated with
520 treatment outcome. *Gut* **67**, 634-643 (2018).
521 14. Kazantseva, J., Malv, E., Kaleda, A., Kallastu, A. & Meikas, A. Optimisation of sample storage and DNA
522 extraction for human gut microbiota studies. *BMC Microbiol* **21**, 158 (2021).
523 15. Maghini, D.G. et al. Quantifying bias introduced by sample collection in relative and absolute microbiome
524 measurements. *Nat Biotechnol* **42**, 328-338 (2024).
525 16. Zhai, X. et al. The impact of storage buffer and storage conditions on fecal samples for bacteriophage
526 infectivity and metavirome analyses. *Microbiome* **11**, 193 (2023).
527 17. Shah, S.A. et al. Expanding known viral diversity in the healthy infant gut. *Nat Microbiol* **8**, 986-998 (2023).
528 18. Zolfo, M., Vol. Ph.D. University of Trento, Trento; 2020.
529 19. Unterer, M., Khan Mirzaei, M. & Deng, L. Gut Phage Database: phage mining in the cave of wonders. *Signal*
530 *Transduct Target Ther* **6**, 193 (2021).
531 20. Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F. & Gordon, J.I. Going viral: next-generation
532 sequencing applied to phage populations in the human gut. *Nat Rev Microbiol* **10**, 607-617 (2012).
533 21. Khan Mirzaei, M. et al. Challenges of Studying the Human Virome - Relevant Emerging Technologies. *Trends*
534 *Microbiol* **29**, 171-181 (2021).
535 22. Marine, R. et al. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput
536 sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol* **77**, 8071-8079 (2011).
537 23. Kim, K.H. & Bae, J.W. Amplification methods bias metagenomic libraries of uncultured single-stranded and

538 double-stranded DNA viruses. *Appl Environ Microbiol* **77**, 7663-7668 (2011).
539 24. Roux, S. et al. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses.
540 *PeerJ* **4**, e2777 (2016).
541 25. Chen, M. et al. Comparison of multiple displacement amplification (MDA) and multiple annealing and
542 looping-based amplification cycles (MALBAC) in single-cell sequencing. *PLoS One* **9**, e114520 (2014).
543 26. Callanan, J. et al. Expansion of known ssRNA phage genomes: From tens to over a thousand. *Sci Adv* **6**,
544 eaay5981 (2020).
545 27. Starr, E.P., Nuccio, E.E., Pett-Ridge, J., Banfield, J.F. & Firestone, M.K. Metatranscriptomic reconstruction
546 reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc Natl Acad Sci U S A* **116**,
547 25900-25908 (2019).
548 28. Sachsenroder, J. et al. Simultaneous Identification of DNA and RNA Viruses Present in Pig Faeces Using
549 Process-Controlled Deep Sequencing. *Plos One* **7** (2012).
550 29. Krishnamurthy, S.R., Janowski, A.B., Zhao, G., Barouch, D. & Wang, D. Hyperexpansion of RNA
551 Bacteriophage Diversity. *PLoS Biol* **14**, e1002409 (2016).
552 30. Neri, U. et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**,
553 4023-4037 e4018 (2022).
554 31. Zhang, T. et al. RNA viral community in human feces: Prevalence of plant pathogenic viruses. *Plos Biology* **4**,
555 108-118 (2006).
556 32. Jurasz, H., Pawłowski, T. & Perlejewski, K. Contamination Issue in Viral Metagenomics: Problems, Solutions,
557 and Clinical Perspectives. *Front microbiol* **12** (2021).
558 33. Martin, S.A.M., Dehler, C.E. & Krol, E. Transcriptomic responses in the fish intestine. *Dev Comp Immunol* **64**,
559 103-117 (2016).
560 34. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98
561 (2011).
562 35. Troll, C.J. et al. A ligation-based single-stranded library preparation method to analyze cell-free DNA and
563 synthetic oligos. *BMC Genomics* **20**, 1023 (2019).
564 36. Wilson, G.G., Young, K.K.Y., Edlin, G.J. & Konigsberg, W. High-Frequency Generalized Transduction by
565 Bacteriophage-T4. *Nature* **280**, 80-82 (1979).
566 37. Tao, P., Wu, X.R., Tang, W.C., Zhu, J.G. & Rao, V. Engineering of Bacteriophage T4 Genome Using
567 CRISPR-Cas9. *Acs Synth Biol* **6**, 1952-1961 (2017).
568 38. Hua, J. in University of Pittsburgh, Vol. Doctoral Dissertation (University of Pittsburgh, University of
569 Pittsburgh; 2016).
570 39. Guerin, E. & Hill, C. Shining Light on Human Gut Bacteriophages. *Front Cell Infect Microbiol* **10**, 481 (2020).
571 40. Jakociune, D. & Moodley, A. A Rapid Bacteriophage DNA Extraction Method. *Method Protocol* **1** (2018).
572 41. Soleimani-Delfan, A., Bouzari, M. & Wang, R. A rapid competitive method for bacteriophage genomic DNA
573 extraction. *J Virol Methods* **293** (2021).
574 42. Teng, Y. et al. Gut bacterial isoamylamine promotes age-related cognitive dysfunction by promoting
575 microglial cell death. *Cell Host Microbe* **30**, 944- 960.e8 (2022).
576 43. Maud, B. et al. An alternative method to multiple displacement amplification for preparing virome DNA in a
577 way adapted for the sequencing of both double-strand and single-strand DNA viruses. *bioRxiv*,
578 2022.2012.2012.520144 (2022).
579 44. Singer, V.L., Jones, L.J., Yue, S.T. & Haugland, R.P. Characterization of PicoGreen reagent and development of
580 a fluorescence-based solution assay for double-stranded DNA quantitation. *Anal Biochem* **249**, 228-238
581 (1997).

582 45. Cui, L. et al. Four additional natural 7-deazaguanine derivatives in phages and how to make them. *Nucleic*
583 *Acids Research* **51**, 9214-9226 (2023).
584 46. Nakayama, Y., Yamaguchi, H., Einaga, N. & Esumi, M. Pitfalls of DNA Quantification Using DNA-Binding
585 Fluorescent Dyes and Suggested Solutions. *Plos One* **11** (2016).
586 47. Gregory, A.C. et al. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the
587 Human Gut. *Cell Host Microbe* **28**, 724-740 e728 (2020).
588 48. Camargo, A.P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework
589 of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research* **51**, D733-D743 (2023).
590 49. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat*
591 *Microbiol* **6**, 960-970 (2021).
592 50. Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D. & Lawley, T.D. Massive expansion of
593 human gut bacteriophage diversity. *Cell* **184**, 1098- 1098-1109.e9 (2021).
594 51. Sato, M.P. et al. Comparison of the sequencing bias of currently available library preparation kits for
595 Illumina sequencing of bacterial genomes and metagenomes. *DNA Res* **26**, 391-398 (2019).
596 52. Jones, M.B. et al. Library preparation methodology can influence genomic and functional predictions in
597 human microbiome research. *Proc Natl Acad Sci U S A* **112**, 14024-14029 (2015).
598 53. Wilcox, A.H., Delwart, E. & Díaz-Muñoz, S.L. Next-generation sequencing of dsRNA is greatly improved by
599 treatment with the inexpensive denaturing reagent DMSO. *Microbial genomics* **5**, e000315 (2019).
600 54. Robertson, G. et al. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**, 909-912 (2010).
601 55. Hoang, M.L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using
602 massively parallel sequencing. *Proc Natl Acad Sci U S A* **113**, 9846-9851 (2016).
603 56. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag, New York; 2016).
604 57. McMurdie, P.J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of
605 Microbiome Census Data. *Plos One* **8** (2013).
606 58. Oksanen, F.J., et al. Vegan: Community Ecology Package. *R package Version 2.6-2* (2022).
607 59. Ahlmann-Eltze, C. & Patil, I. ggsignif: R Package for Displaying Significance Brackets for 'ggplot2'. (PsyArXiv,
608 2021).
609 60. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data
610 with DESeq2. *Genome Biology* **15** (2014).
611 61. Gu, Z.G., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional
612 genomic data. *Bioinformatics* **32**, 2847-2849 (2016).
613 62. Zulkower, V. & Rosser, S. DNA Features Viewer: a sequence annotation formatting and plotting library for
614 Python. *Bioinformatics* **36**, 4350-4352 (2020).
615 63. Thomas, H.M., J. A. ; Kristina, H. gggenomes: A Grammar of Graphics for Comparative Genomics. *R package*
616 *version 0.9.5.9000* (2022).

617 **Table 1** Overview of phage genome characteristics included in the mock communities from the present
618 study

Phage	Family (Species)*	Structure	Genomic type	Genomic Topology	Number of genomic segment(s)	Genomic length (bp)	GC%
C2	**(Ceduovirus c2)	Non-enveloped	dsDNA	Linear	1	22172	36.31%
T4 (T4-c) #	Straboviridae (Tequatrovirus T4)	Non-enveloped	dsDNA	Linear	1	168903	35.30%
Phi29	Salasmaviridae (Salasvirus phi29)	Non-enveloped	dsDNA	Linear	1	19282	40.00%
P1	**(Punavirus P1)	Non-enveloped	dsDNA	Linear	2	94800	47.31%
T7	Autographiviridae (Teseptimavirus T7)	Non-enveloped	dsDNA	Linear	1	39937	48.40%
P35	(Listeria phage P35)	Non-enveloped	dsDNA	Linear	1	35822	40.82%
Lambda	**(Lambdavirus lambda)	Non-enveloped	dsDNA	Linear	1	48502	49.86%
PhiX174	Microviridae (Sinsheimervirus phiX174)	Non-enveloped	**ssDNA (>85%)**	**Circular**	1	5386	44.76%
M13mp18	Inoviridae (Inovirus M13)	Non-enveloped	**ssDNA**	**Circular**	1	7249	42.27%
MS2	Fiersviridae (Emesvirus MS2)	Non-enveloped	**ssRNA**	Linear	1	3569	52.12%
Phi6	Cystoviridae (Cystovirus phi6)	**Enveloped**	**dsRNA**	Linear	3	13385	55.82%

619 * Current taxonomy of ICTV217.
620 **- No family levels.
621 # T4-c has the same genome characteristic as T4, but with heavily modified groups.

622 **Figure Legends**

623 **Fig.1 | Overview of study workflow**. The reproposed library preparation method (SSLR) was indicated
624 with 7 steps (detailed information can be found from **Fig.S1** and **Supplementary file1**). Five different
625 libraries were used to prepare and sequence three artificial bacteriophage mocks containing different
626 proportions of the ssDNA phages (phiX174 and M13mp18) mixed with the dsDNA phages. These phage
627 genome abundance values were calculated based on the quantity of dsDNA and ssDNA phages measured
628 Qubit dsDNA (or ssDNA) HS Assay kit. MA: Mock A with a ratio of ~90:10 for dsDNA and ssDNA (**Fig.S1E**);
629 MB: Mock B with a ratio of ~50:50 for dsDNA and ssDNA; MC: Mock C with a ratio of ~10:90 for dsDNA
630 and ssDNA. MD: Mock D with a ratio of ~90:10 for DNA and RNA; ME: Mock E with a ratio of ~50:50 for
631 DNA and RNA; MF: Mock F with a ratio of 10:90 for DNA and RNA (**Fig.S1F**). MG: Mock G contains high
632 modification T4 genome (T4) with equal ratio of all genomes; MH: Mock H contains lower modification T4
633 genome (T4-c) with equal ratio of all genomes (**Fig.S1G**).

634

635 **Fig.2 | Comparison of different library strategies for DNA mock communities.** (A) Percentage of phage
636 genomes with T4 generated by different library preparation methods. Different color indicate different
637 phage genomes. The percentage of phage genomes was calculated by aligning the clean reads to
638 customed database (**Supplementary file6**) with bowtie2 to assess the relative abundance of each DNA
639 phage (**TableS1**). (B) Principal coordinates analysis (PCoA) plots of bray-Curtis distance matrices. PCoA was
640 used to plot the beta diversity of mock-associated communities using the bray matrix. Different colors
641 indicate different library preparation methods, different shapes indate different DNA mock communites.
642 The dark-filled shapes display the mock with T4 genome and the non-filled without T4. For each axis, in
643 square brackets, the percent of variation explained was reported. (C) Percentage of phage genomes
644 without T4. (D) Pearson correlation coefficient (r) and two-tailed p-value between the expected and
645 obtained read distributions (in percentage) in different DNAs phage mock communities with (top panel)
646 or without (bottom panel) phage T4 genome. Five different libraries were used to prepare and sequence
647 three artificial bacteriophage mocks containing different proportions of the ssDNA phage phiX174 and
648 M13mp18 mixed with the dsDNA phage. These phage genome abundance values were calculated based
649 on the quantity of dsDNA and ssDNA phages measured Qubit dsDNA (or ssDNA) HS Assay kit. MA: Mock A
650 with a ratio of ~90:10 for dsDNA and ssDNA; MB: Mock B with a ratio of ~50:50 for dsDNA and ssDNA; MC:
651 Mock C with a ratio of ~10:90 for dsDNA and ssDNA.

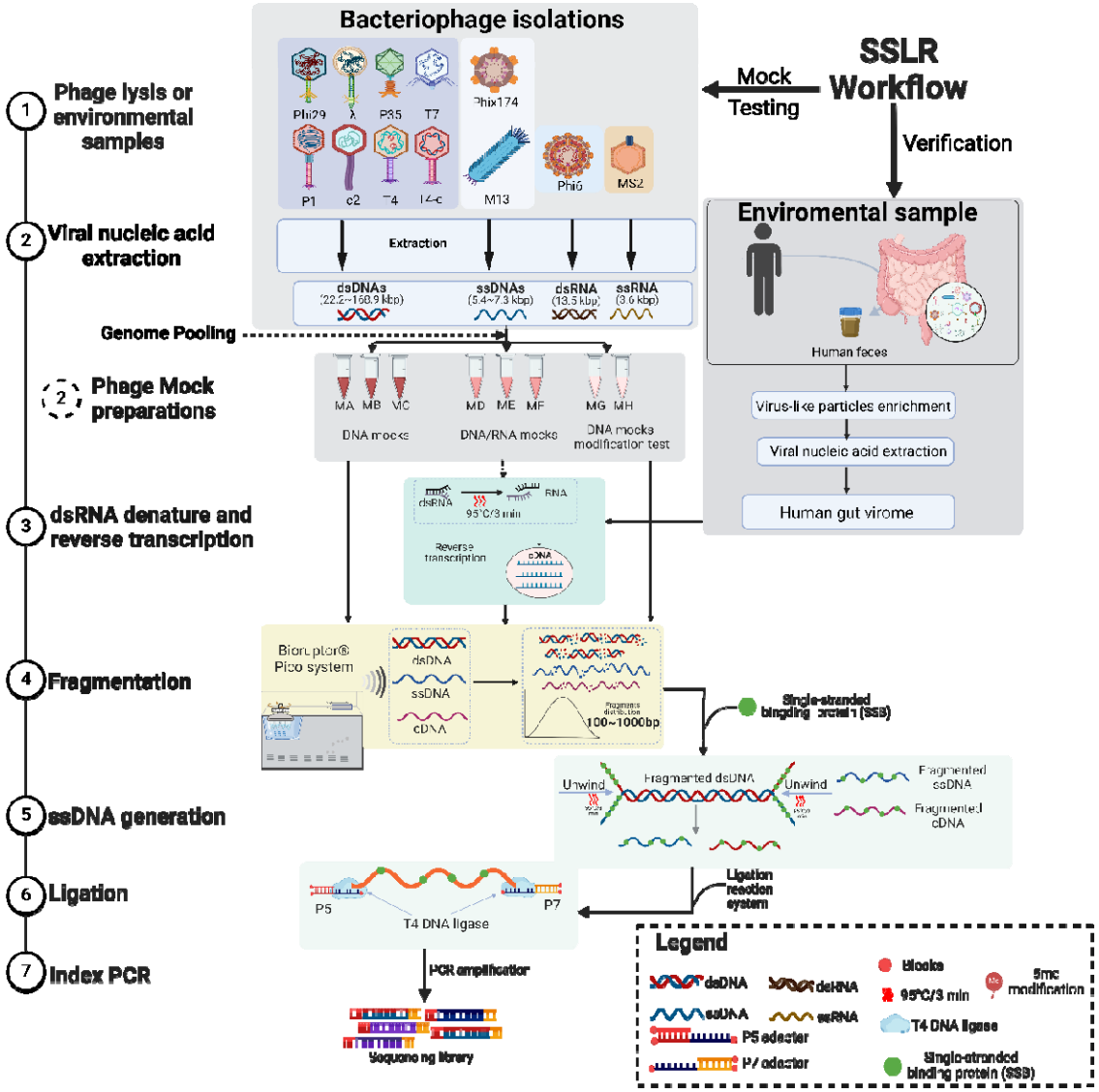
652

653 **Fig.3 | Quantification and MDA bias for highly modified T4 genome**. (A) The concentrations of DNAs
654 used in this study were measured by two different quantification methods (Qubit and Nanodrop). (B)
655 MDA amplification differ in genomes. One and 10 ng of DNA from P1, T4, T4-c and T7 were amplified by
656 30 min of MDA and purified with Zymol Genomic Purification kit. The y-axis indicates the total yield of
657 amplified MDA products (μg)and measured either by Nanodrop or Qubit. (C) The genomic DNA (T4, T4-c,
658 P1 and T7) and the MDA amplification products (2 repeats for each reaction) were visualized by
659 Tapestation4200 with genomic screentapes. (D) Percentage of phage genomes with T4 or T4-c generated
660 by different library preparation methods. Different color indicate different phage genomes with equally
661 input (11.11%). MG: Mock G contains high modification T4 genome (T4) with equal ratio of all genomes;
662 MH: Mock H contains lower modification T4 genome (T4-c) with equal ratio of all genomes (**Fig.S1G**). The
663 percentage of phage genomes was calculated by aligning the clean reads to customed database
664 (**Supplementary file6**) with bowtie2 to assess the relative abundance of each DNA phage (**TableS5**).

665

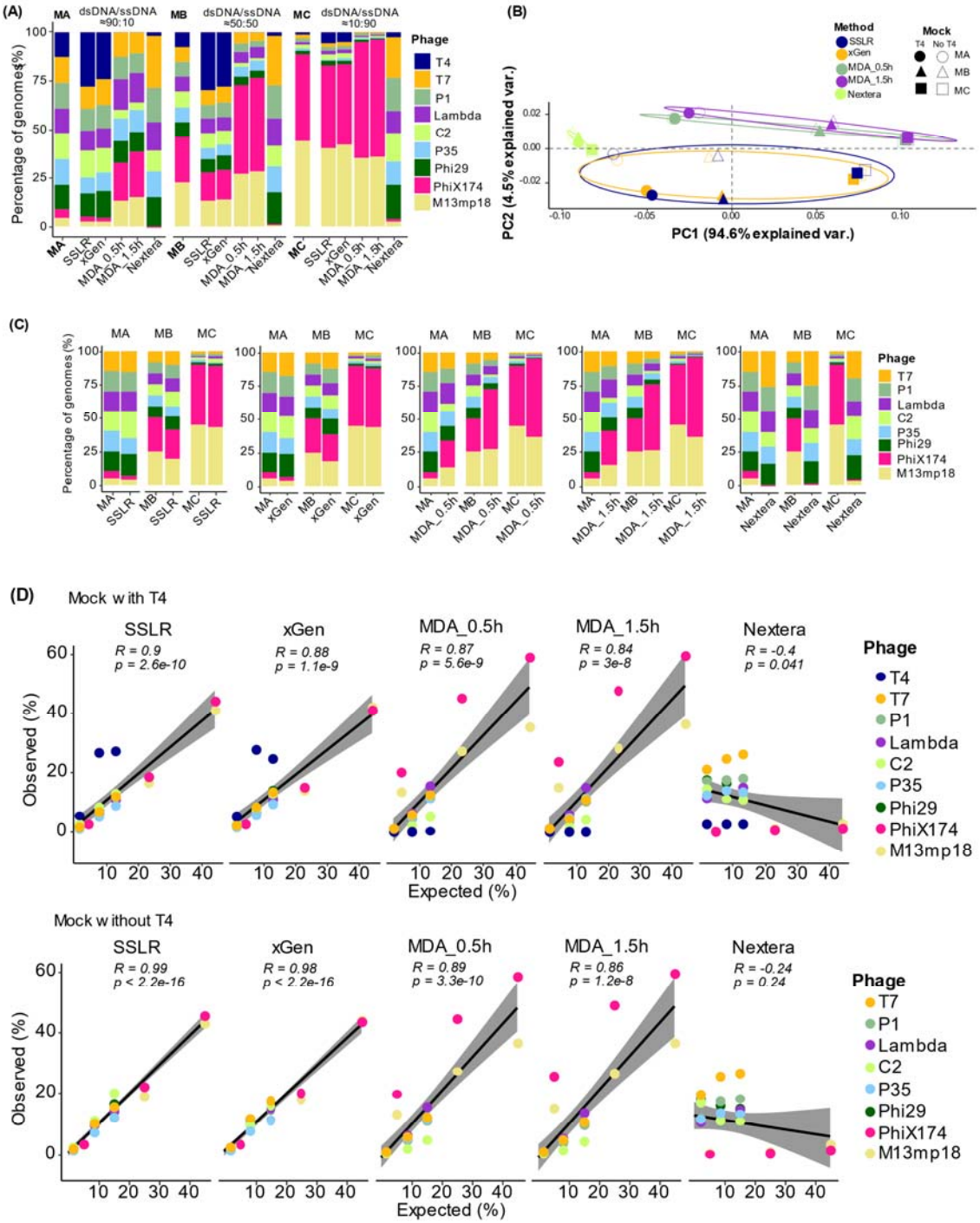
666 **Fig.4 | Comparison of the efficiency of SSLR on the simultaneous identification of DNA/RNA mock**
667 **communities**. SSLR was used to prepare and sequence 4 artificial virome containing different proportions
668 of the DNA phage and RNA phage. (A) Effect of DMSO and heat treatment on the percentage of 4 phage
669 genomes. The percentage of phage genomes was calculated by aligning the clean reads to customed
670 database (**Supplementary file6**) with bowtie2 to assess the relative abundance of each DNA phage
671 (**TableS6**). These phage genome abundance values were calculated based on the quantity of DNA and
672 RNA phages measured by Nanodrop. MD: Mock D with ratio of 90:10 for DNA and RNA; ME: Mock E with
673 ratio of 50:50 for DNA and RNA; MF: Mock F with ratio of 10:90 for DNA and RNA. phi6 has 3 segments,
674 large (6,374 bp), medium (4,063 bp) and small (2,948 bp). (B) Heat treatment does not adversely affect
675 sequencing error rates. The R package ShadowRegression estimates reference-free error rates (inset)
676 based on a transform of the slope of read counts and their 'shadows' (main plot line graphs). Shadows
677 (y-axis) are a measure of the variation in read counts across different sequencing runs for the same
678 sample. They are calculated by taking the logarithm of the ratio of read counts in one run to another run.
679 A higher shadow value means a larger difference in read counts between the two runs. Tags (x-axis) are a
680 measure of the abundance of reads for a given nucleotide position in a sample. They are calculated by
681 taking the logarithm of the read count at that position. A higher tag value means a higher number of
682 reads at that position. The figure shows the relationship between shadows and tags for different samples
683 treated with or without Heat/DMSO. The slope of this relationship is used to estimate the sequencing
684 error rate for each sample, which is shown in the inset plots. The figures suggest that DMSO treatment
685 does not affect the sequencing error rate significantly.
686
687 **Fig.5 | Comparison of different library strategies for fecal virome communities.** (A) Principal coordinates
688 analysis (PCoA) plots of Bray-Curtis distance matrices. PCoA was used to plot the beta diversity of
689 viral-associated communities using the bray matrix. Different colors indicate different library preparation
690 methods. For each axis, in square brackets, the percentage of variation explained was reported. (B)
691 Representative taxonomic distribution (relative abundance) of the sequenced virome. The relative
692 distribution is described at the taxonomical level of the family. The taxonomy of contigs was determined
693 by querying the viral contigs against a database containing taxon signature genes for virus orthologous
694 group hosted at www.vogdb.org. The unclassified are the contigs that cannot be assigned to any known
695 viral taxonomy at the family level, the unknown is the contigs that are related to "viral dark matter". (C)
696 Volcano plot illustrates differential vOTUs were derived from DESeq2 analysis that had more than or equal
697 to two-fold changes in SSLR versus xGen. Each dot represents a vOTU contig and is colored to indicate
698 significance. Grey: not significant (NS), Green – significant by log2 fold change (>2), Blue: significant by
699 p-value (<0.01), Red – significant by log2 fold change and p-value. Figure made with EnhancedVolcano
700 implemented in R. (D) Heatmap of the high abundance (RPKM) viral contigs that were highly enriched in
701 SSLR method that does not present in the rest of methods. This figure shows a subset of the data
702 presented in **Fig.S8A**. Genomic maps of the open reading frames (ORFs) which are predicted by prodigal
703 and then annotated by the blast from the NCBI protein database, the best hits were used for the
704 annotation. Different colors indicate different annotated protein, directional boxes indicate ORFs in the
705 respective orientation (detailed contigs legend can be found from **Fig.S9D**.

706 **Fig.1**

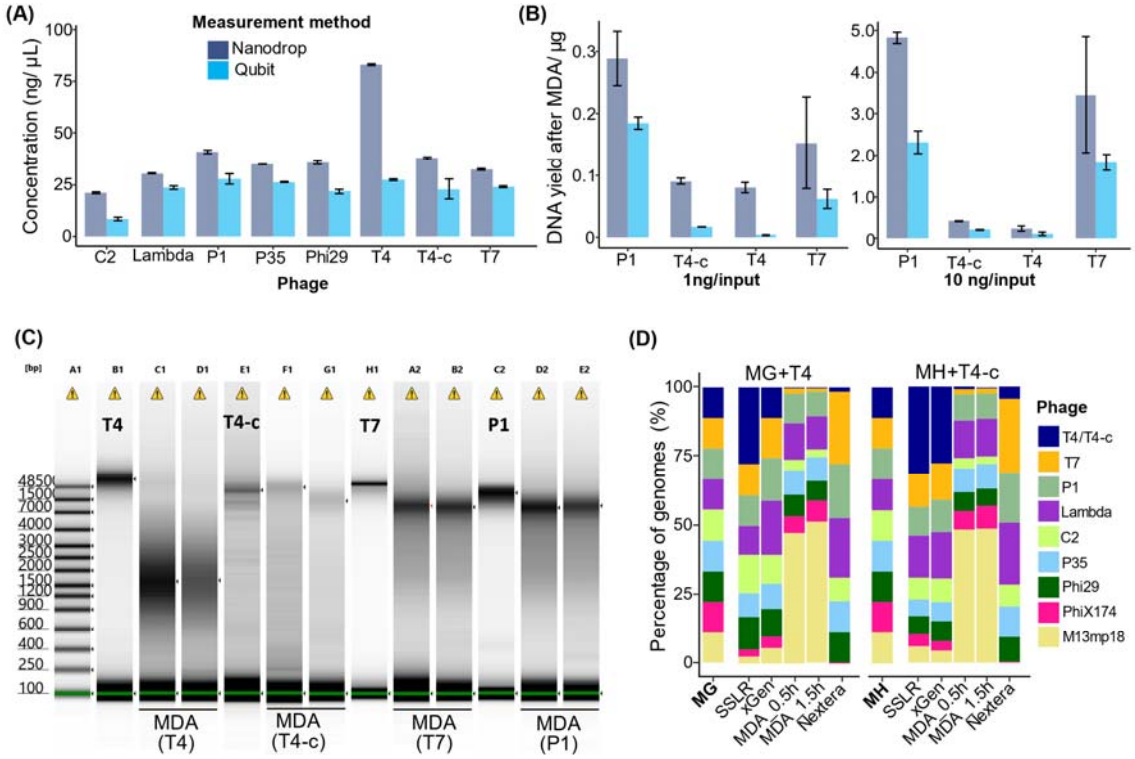


707
708

Fig.2


(A)
MA
dsDNA/ssDNA ≈90:10
MB
dsDNA/ssDNA ≈50:50
MC
dsDNA/ssDNA ≈10:90
Percentage of genomes (%)
Phage
T4
T7
P1
Lambda
C2
P35
Phi29
PhiX174
M13mp18
SSLR
xGen
MDA_0.5h
MDA_1.5h
Nextera
(B)
Method
SSLR
xGen
MDA_0.5h
MDA_1.5h
Nextera
Mock
T4
No T4
MA
MB
MC
PC2 (4.5% explained var.)
PC1 (94.6% explained var.)
(C)
Percentage of genomes (%)
Phage
T7
P1
Lambda
C2
P35
Phi29
PhiX174
M13mp18
(D)
Mock with T4
SSLR
R = 0.9
p = 2.6e-10
xGen
R = 0.88
p = 1.1e-9
MDA_0.5h
R = 0.87
p = 5.6e-9
MDA_1.5h
R = 0.84
p = 3e-8
Nextera
R = -0.4
p = 0.041
Observed (%)
Expected (%)
Phage
T4
T7
P1
Lambda
C2
P35
Phi29
PhiX174
M13mp18
Mock without T4
SSLR
R = 0.99
p < 2.2e-16
xGen
R = 0.98
p < 2.2e-16
MDA_0.5h
R = 0.89
p = 3.3e-10
MDA_1.5h
R = 0.86
p = 1.2e-8
Nextera
R = -0.24
p = 0.24
Observed (%)
Expected (%)
Phage
T7
P1
Lambda
C2
P35
Phi29
PhiX174
M13mp18

713 **Fig.3**



714

715

Fig.4

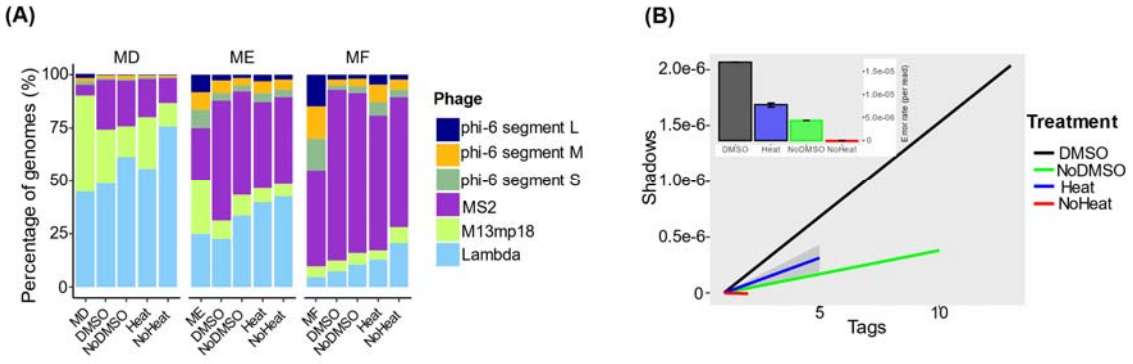
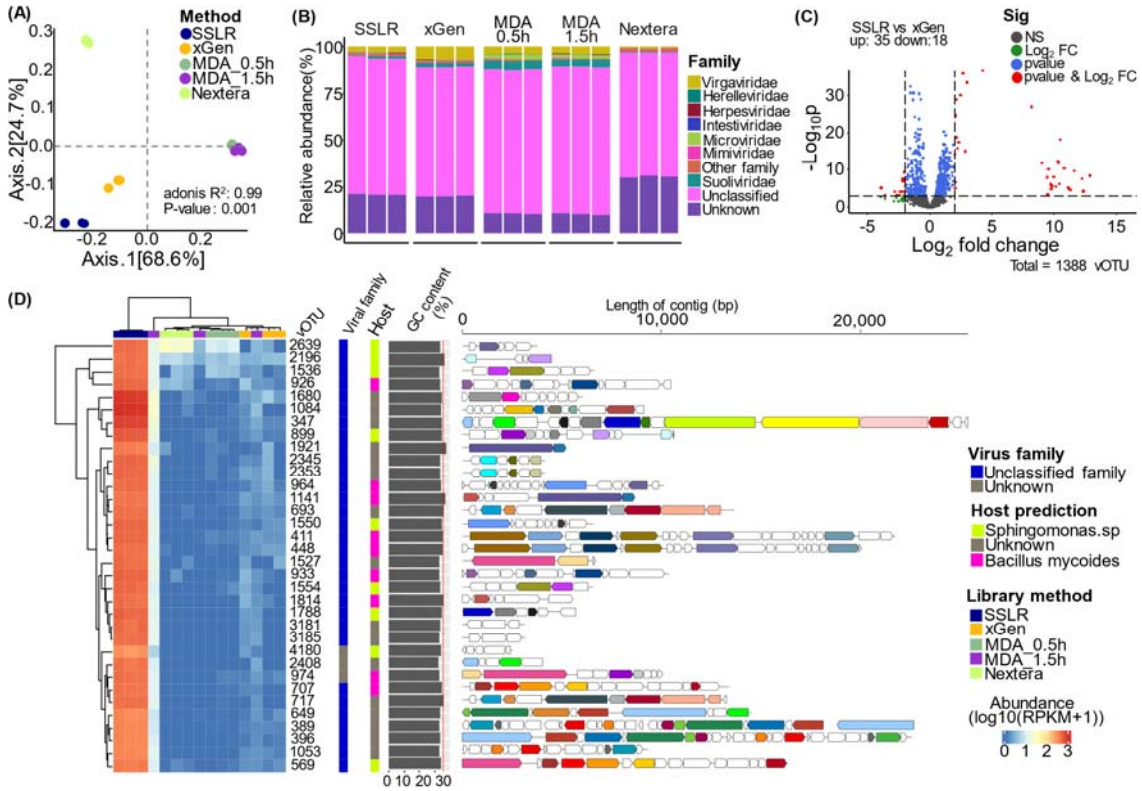

(A)
MD
ME
MF
Percentage of genomes (%)
100
75
50
25
0
MD
DMSO
NoDMSO
Heat
NoHeat
Phage
phi-6 segment L
phi-6 segment M
phi-6 segment S
MS2
M13mp18
Lambda
(B)
Shadows
2.0e-6
1.5e-6
1.0e-6
0.5e-6
0
5
10
Tags
Treatment
DMSO
NoDMSO
Heat
NoHeat

Fig.5


(A)
Method
SSLR
xGen
MDA_0.5h
MDA_1.5h
Nextera
Axis.2[24.7%]
Axis.1[68.6%]
adonis R²: 0.99
P-value: 0.001
(B)
SSLR
xGen
MDA 0.5h
MDA 1.5h
Nextera
Relative abundance(%)
Family
Virgaviridae
Herelleviridae
Herpesviridae
Intestiviridae
Microviridae
Mimiviridae
Other family
Suoliviridae
Unclassified
Unknown
(C)
SSLR vs xGen
up: 35 down:18
Sig
NS
Log2 FC
pvalue
pvalue & Log2 FC
-Log10p
Log2 fold change
Total = 1388 vOTU
(D)
vOTU
Viral family
Host
GC content (%)
Length of contig (bp)
Virus family
Unclassified family
Unknown
Host prediction
Sphingomonas.sp
Unknown
Bacillus mycoides
Library method
SSLR
xGen
MDA_0.5h
MDA_1.5h
Nextera
Abundance (log10(RPKM+1))